NATURAL LANGUAGE COMPUTING
Natural language

• Humans have been recording our languages for much of our $\geq 100,000$ year existence. E.g.,
  • Poems and songs.
  • Speeches and lectures.
  • News broadcasts.
  • Short stories, novels, and articles.
  • Essays and reports for school and work.
  • Emails and text messages.
  • Signs and graffiti on buildings, streets, and people.

• We may owe our existence (and certainly our civilization) to our ability to **produce** and **understand** language.
What is natural language computing?

Getting computers to understand everything we say and write.

In this class (and in the field generally), we are interested in the **statistics of language**.

(Occasionally, computer models give insight into how humans process language.)
What can natural language do?

The ultimate in human-computer interaction.

“translate Also Sprach Zarathustra”
“take a memo…”
“open the pod bay doors”
“how far until Jupiter?”
“Can you summarize 2001: A Space Odyssey?”

We’re making progress, but why are these things still hard to do?
A little deeper

• Language has **hidden structures**, e.g.,
  • How are **sounds** and **text** related?
    • e.g., why is this: not a ‘ghoti’ (enough, women, nation)?

• How are words **combined** to make sentences?
  • e.g., what makes ‘**colourless green ideas sleep furiously**’ correct in a way unlike ‘**furiously sleep ideas green colourless**’.

• How are words and phrases used to produce **meaning**?
  • e.g., why is ‘**yes**’ an **inappropriate** response to the question ‘**do you know what time it is?**’?

• What do we already know about these hidden structures?
Categories of linguistic knowledge

- **Phonology**: the study of patterns of speech sounds.
  
  e.g., “read” → /r iy d/

- **Morphology**: how words can be changed by inflection or derivation.
  
  e.g., “read”, “reads”, “reader”, “reading”, ...

- **Syntax**: the ordering and structure between words and phrases (i.e., grammar).
  
  e.g., \( \text{NounPhrase} \rightarrow \text{article adjective noun} \)

- **Semantics**: the study of how meaning is created by words and phrases.
  
  e.g., “book” → 

- **Pragmatics**: the study of meaning in contexts.
Ambiguity – Phonological

• **Phonology**: the study of patterns of speech sounds.

<table>
<thead>
<tr>
<th>Word</th>
<th>Phonemes</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;read&quot;</td>
<td>/r iy d/</td>
<td>as in ‘I like to read’</td>
</tr>
<tr>
<td>&quot;read&quot;</td>
<td>/r eh d/</td>
<td>as in ‘She read a book’</td>
</tr>
<tr>
<td>&quot;object&quot;</td>
<td>/aa b jh eh k t/</td>
<td>as in ‘That is an object’</td>
</tr>
<tr>
<td>&quot;object&quot;</td>
<td>/ah b jh eh k t/</td>
<td>as in ‘I object!’</td>
</tr>
<tr>
<td>&quot;too&quot;</td>
<td>/t uw/</td>
<td>as in ‘too evil’</td>
</tr>
<tr>
<td>&quot;two&quot;</td>
<td>/t uw/</td>
<td>as in ‘two beers’</td>
</tr>
</tbody>
</table>

• Ambiguities can often be **resolved** in context, but not always.
  • e.g., /h aw t uw r eh k ah ?? n ay z s (b|p) iy ch/ → ‘how to recognize speech’
    → ‘how to wreck a nice beach’
Resolution with syntax

- If you hear the sequence of speech sounds
  
  `/ay sh aa t ae n eh l ah f ah n t ih n m ay p ih jh ae m ae z/

  which word sequence is being spoken?
  
  → “Eye Shaw tan L F aunt inn my pee jam as”
  → “I Shaw tan L F aunt inn my pee jam as”
  → “Eye shot an L laugh an tin my P jam as”
  → “I shot an L laugh an tin my P jam as”
  → “I shot an elephant inn my pee jam as”
  → “I shot an elephant in my pyjamas”

- It’s obvious (to us) that the last option is most likely because we have knowledge of syntax, i.e., grammar.
Ambiguity – Syntactic

- **Syntax**: the ordering and structure between words. Words can be grouped into ‘parse tree’ structures given grammatical ‘rules’.

  e.g., “I shot an elephant in my pyjamas”
Resolution with semantics

- It’s obvious (to us) that the elephants don’t wear pyjamas, and we can discount one option because of our knowledge of semantics, i.e., meaning.
Aside – ‘garden path’ sentences

A garden path sentence is one in which you are led to assume one syntactic structure before being forced into another.

- The man who hunts ducks out on weekends.
- The cotton clothing is made of grows in Mississippi.
- Fat people eat accumulates.
- Buffalo buffalo Buffalo buffalo buffalo Buffalo buffalo.
- The horse raced past the barn fell.
- Mary gave the child the dog bit a bandaid.
- All women who admire a man who paints like Monet.
Ambiguity – Semantic

- **Semantics**: the study of how meaning is created by the use of words and phrases.

  - “Every man loves a woman”
    - \( \forall x \, \text{man}(x) \exists y: (\text{woman}(y) \land \text{loves}(x, y)) \)
    - \( \exists y: \text{woman}(y) \land \forall x (\text{man}(x) \rightarrow \text{loves}(x, y)) \)

  - “I made her duck”
    - I cooked waterfowl meat for her to eat.
    - I cooked waterfowl that belonged to her.
    - I carved the wooden duck that she owns.
    - I caused her to quickly lower her head.

  - “Give me the pot”
    - It’s time to bake.
    - It’s time to get baked.
Resolution with pragmatics

- It’s obvious (to us) which meaning is intended given knowledge of the context of the conversation or the world in which it takes place.

  - “Every man loves a woman”
    \[
    \forall x \text{ man}(x) \exists y: (\text{woman}(y) \land \text{loves}(x, y)) \\
    \exists y: \text{woman}(y) \land \forall x (\text{man}(x) \rightarrow \text{loves}(x, y))
    \]

  - “I made her duck”
    \[
    \rightarrow \text{I cooked waterfowl meat for her to eat.} \\
    \rightarrow \text{I cooked waterfowl that belonged to her.} \\
    \rightarrow \text{I carved the wooden duck that she owns.} \\
    \rightarrow \text{I caused her to quickly lower her head.}
    \]

  - “Give me the pot”
    \[
    \rightarrow \text{It’s time to bake.} \\
    \rightarrow \text{It’s time to get baked.}
    \]
Ambiguity – miscellaneous

- Newspaper headlines (spurious or otherwise)

  - Kicking Baby Considered to be Healthy
  - Squad Helps Dog Bite Victim
  - Canadian Pushes Bottle Up Germans
  - Milk Drinkers are Turning to Powder
  - Grandmother of Eight Makes Hole in One
  - Kids Make Nutritious Snacks
  - Juvenile Court Tries Shooting Defendant
  - Local High School Dropouts Cut in Half

CSC401/2511 – Spring 2014
Q: Ambiguity

- What kind of ambiguity is found here?
  - Phonological, syntactic, or semantic?

Harper Wins on Budget, More Lies Ahead

Teacher Strikes Idle Kids

Police: Crack found in defendant’s buttocks

Syntactic: Is ‘Lies’ a noun or a verb?

Syntactic: Is ‘Strikes’ a noun or a verb?

Semantic: What is the meaning of ‘crack’?
NLC as Artificial Intelligence

• NLC involves **resolving ambiguity** at all levels.
  • **Reasoning with world knowledge.**
    • In the early days knowledge was **explicitly encoded** in artificial **symbolic** systems (e.g., context-free grammars) by **experts**.
  
• Now, algorithms learn using **probabilities** to distinguish subtly different competing hypotheses.
  • E.g., is *Google* a **noun** or a **verb**?
  • An example where *Google* \( \epsilon \) **Nouns** ("*Google makes Android*"), does **not** mean that Google is never a verb ("*I Googled your mom*”).

\[
P(\text{Google } \epsilon \text{ Nouns}) > P(\text{Google } \epsilon \text{ Verbs}) > 0
\]
The Turing Test

• First and most fundamental test of machine intelligence.

• A machine (A) imitates a human using nothing but a text-based instant messenger.
  • If a human interrogator (C) cannot reliably differentiate a real human (B) from the machine, that machine is said to be ‘intelligent’.

Aside – Chatterbots

• ELIZA (Weizenbaum, 1966): simple pattern matching to imitate a psychiatrist.

• Surprisingly effective despite no linguistic knowledge.

• e.g.,

  User: Men are all alike.
  ELIZA: In what way?
  User: They’re always bugging us about something or other.
  ELIZA: Can you think of a specific example?
  User: My boyfriend made me come here.
  ELIZA: Your boyfriend made you come here.

  (Jurafsky and Martin, 2009)

http://www.cleverbot.com
Course outline (approximate)

• Introduction and linguistic data (2 lectures)
• $N$-gram models and features of data (2 lectures) *
• Entropy and information theory (2 lectures) *
• Hidden Markov models (3 lectures) *
• Statistical machine translation (3 lectures) **
• Articulatory and acoustic phonetics (2 lectures) *
• Automatic speech recognition (2 lectures) **
• Speech synthesis (1 lecture) **
• Information retrieval (2 lectures) **
• Text summarization (1 lecture) **
• Other classifiers and review (2 lectures)

* techniques  ** applications
Preview: Machine translation

- One of the most prized applications in NLC.
- Requires both **interpretation** and **generation**.
- Over $50B spent annually on human translation.
Human: According to the data provided today by the Ministry of Foreign Trade and Economic Cooperation, as of November this year, China has actually utilized 46.959B US dollars of foreign capital, including 40.007B US dollars of direct investment from foreign businessmen.

IBM4: The Ministry of Foreign Trade and Economic Cooperation, including foreign direct investment 40.007B US dollars today provide data include that year to November China actually using foreign 46.959B US dollars and

Yamada/Knight: Today’s available data of the Ministry of Foreign Trade and Economic Cooperation shows that China’s actual utilization of November this year will include 40.007B US dollars for the foreign direct investment among 46.959B US dollars in foreign capital.
Preview: Machine translation

- In the 1950s and 1960s direct word-for-word replacement was popular.
- Due to semantic and syntactic ambiguities and differences in source languages, results were mixed.

"The spirit is willing, but the flesh is weak"

"The vodka is good, but the meat is rotten"

US English

Russian
Preview: Machine translation

• One problem is disparity of meanings in languages.

**nation** *n.* a large body of people, associated with a particular **territory**, that is sufficiently conscious of its **unity** to seek or to possess a **government** of its own

**nation** *n.* an aggregation of persons of the same **ethnic family**, often speaking the same **language** or cognate **languages**
Preview: Machine translation

- **Solution**: automatically learn statistics on parallel texts

  ... citizen of Canada has the right to vote in an election of members of the House of Commons or of a legislative assembly and to be qualified for membership ...

  e.g., the *Canadian Hansards*: bilingual Parliamentary proceedings

  ... citoyen canadien a le droit de vote et est éligible aux élections législatives fédérales ou provinciales ...
Statistical machine translation

- Modern statistical machine translation is based on the following perspective...

When I look at an article in Russian, I say: ‘This is really written in English, but it has been **coded** in some strange symbols. I will now proceed to **decode**.’

Warren Weaver
March, 1947

Claude Shannon
July, 1948
Aside – Machine translation

- http://www.translationparty.com uses Google Translate to go back and forth between English and Japanese until we get two consecutive identical English phrases.
Preview: Machine translation

Start with an English phrase:

that's one small step for a man, one giant leap for mankind

find equilibrium

Equilibrium found!
Okay, I get it, you like Translation Party.
Preview: Speech recognition

My hands are in the air.

Buy ticket... AC490... yes

Put this there.

Dictation
Telephony
Multimodal interaction
Speech waveforms

“Two plus seven is less than ten”
Phoneme sequences

- Often, we assume that a spoken utterance can be partitioned into a sequence of non-overlapping phonemes.
  - This approach is fraught with problems (e.g., when exactly does one phoneme end and another begin?), but it’s useful.

- How do we get information about what phoneme is being uttered?

![Waveform](image-url)
Spectrograms

• Speech sounds can be thought of as overlapping sine waves.
• Speech is **split apart** into a 3D graph called a ‘spectrogram’.
• Spectrograms allow machines to extract **statistical features** that differentiates between different kinds of sounds.
Speech recognition

beet /biːt/  
bat /bæt/  
bott /bat/  
boot /but/
• In order to classify an unknown observation (e.g., $X$), we need a statistical model of the distribution of sounds.
Preview: Questions and answers

Which woman has won more than 1 Nobel prize?

(Marie Curie)

- **Question Answering** (QA) and **Information Retrieval** (IR) involve many of the same principles.
Preview: Information retrieval

Google

WolframAlpha

which woman has won more than 1 Nobel prize?

About 503,000 results (0.31 seconds)

- Who has won more than one Nobel prize? - The Times of India
- Answers.com - Who won the Nobel Prize more than once
- Nobel Prize - Wikipedia, the free encyclopedia
- List of Nobel laureates - Wikipedia, the free encyclopedia
- Has anyone ever won Nobel prizes in more than one category over ...

Input interpretation:
Nobel Prize

<table>
<thead>
<tr>
<th>year</th>
<th>recipient</th>
<th>field</th>
<th>country of achievement</th>
<th>country of birth</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>Akira Suzuki</td>
<td>chemistry</td>
<td>Japan</td>
<td>Japan</td>
</tr>
<tr>
<td>2010</td>
<td>Ei-ichi Negishi</td>
<td>chemistry</td>
<td>United States</td>
<td>Japan</td>
</tr>
<tr>
<td>2010</td>
<td>Richard R. Heck</td>
<td>chemistry</td>
<td>United States</td>
<td>United States</td>
</tr>
<tr>
<td>2010</td>
<td>Christopher A. Pissarides</td>
<td>economics</td>
<td>United Kingdom</td>
<td>Cyprus</td>
</tr>
<tr>
<td>2010</td>
<td>Dale T. Mortensen</td>
<td>economics</td>
<td>United States</td>
<td>United States</td>
</tr>
<tr>
<td>2010</td>
<td>Peter A. Diamond</td>
<td>economics</td>
<td>United States</td>
<td>United States</td>
</tr>
<tr>
<td>2010</td>
<td>Mario Vargas Llosa</td>
<td>literature</td>
<td>Peru</td>
<td>Peru</td>
</tr>
</tbody>
</table>
Aside – Question answering

**WolframAlpha**

How much potassium is in 450,000 cubic kilometers of bananas?

Input interpretation:

| banana | amount | 450 000 km$^3$ (cubic kilometers) | potassium |

Result:

$1.5 \times 10^{12}$ t (metric tons)

Weekly Forecast:

- **TUES**: 14°
- **WED**: 16°
- **THU**: 17°
- **FRI**: 17°
- **SAT**: 18°
- **SUN**: 18°
Answer questioning?

- Retrieving information can be a clever combination of many very simple concepts and algorithms.
Automatic summarization

Russia fights Napoleon and a Natalia likes Boris.

Don’t sit on a wall if you’re an egg.

Gregor turns into a bug.

Girl kills a woman, forms a gang, and kills again.
Overview: NLC

• Is natural language computing (the discipline) hard?
  • Yes, because natural language
    • is highly ambiguous at all levels,
    • is complex and subtle,
    • is fuzzy and probabilistic,
    • involves real-world reasoning.
  • No, because computer science
    • gives us many powerful statistical techniques,
    • allows us to break the challenges down into more manageable features.
NLC in industry
Aside – Interpreting brain signals

• How does the brain **produce** and **understand** language?
• Can we **think a sentence** and have a computer write it out on screen?
Natural language computing

- **Instructor**: Frank Rudzicz ([frank@cdf](mailto:frank@cdf))
- **TAs**: Varada Kolhatkar, Aida Nematzadeh, TBD, TBD
- **Meetings**: Mondays/Wednesdays 10h00-11h00 in RW110
- **Languages**: English, Python, Matlab.
- **Website**: [http://www.cs.toronto.edu/~frank/csc401/](http://www.cs.toronto.edu/~frank/csc401/)
- **You**: Understand basic **probability**, can **program**, or can pick these up as we go.
- **Syllabus**: Key **theory** and **methods** in statistical natural language computing.
  Focus will be on **Markov models, machine translation, and speech recognition**.
Office hours

• **Time:**
  - Immediately after lecture.

• **Location:**
  - BA 4237 until further notice.
  - My real office is at 550 University.
Course information

http://www.cs.toronto.edu/~frank/csc401
Evaluation policies

- **General:** Three assignments: 20% each
  - Final exam: 40%

- **Lateness:** 10% deduction applied to electronic submissions that are 1 minute late.
  - Additional 10% applied every 24 hours up to 72 hours total, at which point grade is zero.

- **Final:** If you fail the final exam, then you fail the course.

- **Ethics:** Plagiarism and unauthorized collaboration can result in a grade of zero on the homework, failure of the course, or suspension from the University.
  - See the course website.
Assignments

- Programming-intensive and individually accomplished.
  - Languages: Python and MATLAB (opt. C/C++ modules).

- Subjects:
  - Assignment 1: Corpus statistics and Twitter classification
    Statistical techniques and classification.
  - Assignment 2: Statistical machine translation
    Statistical $n$-grams, smoothing, and multilingual word alignment.
  - Assignment 3: Automatic speech recognition
    Signal processing, phonetics, and hidden Markov models.
Assignment 1 – Twitter classification

• Involves:
  • Working with social media data (i.e., gathering statistics on some data from Twitter),
  • Part-of-speech tagging (more on this later),
  • Classification with WEKA (e.g., SVM, decision trees).

• You should get an early start.
Projects – graduate students only

• Graduate students can **optionally** undertake a full-term **project** worth **60%** of their grade **instead** of the assignments.
  • Good for those, e.g., who prefer to work in teams. You might even get a publication!

• Teams must consist of 1 or 2 humans (no more, no fewer).
• Projects must contain a significant programming and scientific component.
• Projects must involve **corpus statistics** and at least one of **statistical machine translation** and **automatic speech recognition**.
Projects – graduate students only

• Some possible ideas for projects include:
  • A novel speech-to-speech machine translation.
  • A novel method of using data in language $A$ to train a classification system in language $B$ for $A \neq B$.
  • A system that searches for documents (text and speech) in language $A$ given queries in language $B$ for $A \neq B$.

• If you decide to take this option, you have to notify me by email about your team by 13 January!

• You will need to periodically submit checkpoints that build on their antecedents.
  • See course webpage for detailed requirements!
Reading

Mandatory (and FREE online!)

FOUNDATIONS OF
STATISTICAL NATURAL LANGUAGE
PROCESSING

CHRISTOPHER D. MANNING AND
HINRICH SCHÜTZE

http://cognet.mit.edu/library/books/view?isbn=0262133601

Optional

SPEECH AND LANGUAGE PROCESSING
An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition

DANIEL JURAFSKY & JAMES H. MARTIN
Assignment 1 and reading

• Assignment 1 available now (on course webpage)!
  • Due 7 February
  • TA: Varada Kolhatkar (t9kolhat@cdf)

• Reading:
  • Manning & Schütze: Sections 1.3—1.4.2, Sections 6.0—6.2.1.